..... Fit Hub Gym Center .....

Neeraj Tripathi

20 September 2020

Setting Up a Gym Center in Chicago, Illinois-USA

1. Introduction

1.1 Background

A modern day gymnasium (as gym used to be called way back in Ancient Greece) is a place for indoor physical workout where various equipment and machines are typically used. For some people, a typical gym is a place where you focus on weight lifting and similar activities. Going to a gym has numerous benefits like a great boost in physical and mental health, stress release, confidence in day to day routine work and a place where one can get a change of pace from the normal routine of the day.

In the busy modern life style, health problems are becoming common day by day because of negligence of people towards staying fit. Simple and consistent exercises can keep a lot of health issues at bay. But due to lack of discipline and consistency, and sometimes (though rare) lack of knowledge discourages people from keeping up with a good exercise routine. This behaviour creates the space for organised systems called Gym/ Fitness Centres/ Health Hubs etc. They provide a common place and a perfect environment for workout as they cater to many similar minded people who are conscious about their health. Thus opening gym centres which provide good facilities/services is profitable business consideration for many.

1.2 Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Chicago, Illinois, USA to open a new Gym/Fitness center. Using data science methodology and Machine Learning techniques like clustering, this project aims to provide solutions to answer the business question: What is the best location to open up a new Gym/Fitness center in Chicago?
This project specifically analyses the neighborhoods of Chicago city, Illinois-USA and suggests suitable locations to open up gym centres.

1.3 Stakeholders

This project is useful for business owners/ people who have an interest in opening up a new gym center in Chicago, USA. This is one of the best time to open up fitness centre in this city, as people here are getting increasingly aware towards their health. The health department has also been campaigning their “Healthy Chicago Plan”, so this is the most apt opportunity to grab loyal customers by providing them the quality and type of service they expect from a standard fitness centre.

2. Data

2.1. Source

To approach this business problem, the required data are :

2.1.1. Data about the neighborhoods of Chicago: Name, Zip code, Latitude, Longitude. This data is obtained from two sources:

2.1.1.1. A .csv file containing all the zip codes and corresponding city, coordinates and time zone for each zip code in the State of Illinois, USA. Source: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/download/?format=csv&refine.state=IL

2.1.1.2. Data containing zip codes and neighborhoods of all zip codes for the city of Chicago. Source: https://www.seechicagorealestate.com/chicago-zip-codes-by-neighborhood.php

2.1.2. Data about the places in each of these neighborhoods (venue data): Restauranctes, shops, hotels, bars, clubs, stores, malls, gyms etc. This data is obtained via Foursquare API.

2.2. Description

The data obtained from source mentioned in 2.1.1.1. contains zip code, state, city, latitude, longitude, time zone and daylight savings time flag as columns for all the zip codes that belong to the State of Illinois, USA. The data obtained from source mentioned in 2.1.1.2. contains zip code and neighborhood names for all the zip codes that belong to the city of Chicago. There are multiple neighborhoods for some zip codes as expected, so only the first neighborhood is extracted after scraping this data from the source in order to preserve a unique zip to neighborhood mapping. The data from both these sources are then merged together to obtain the required data for neighborhoods of Chicago. This dataset has 99 neighborhoods in total.

After the initial preparation, the venue data is obtained for each neighborhood present in the dataset using Foursquare API. This returns the name, coordinates and the category venues for each of the neighborhoods.

2.2.1. Data Preprocessing

The unnecessary features were discarded after the merge step of neighborhood data and the zip code data, and the final dataset thus obtained has the following features: Zip, Neighborhood, Latitude, Longitude. Next, the venue data is obtained. This results in 9900 venues belonging to 250 venue categories, 100 venues for each neighborhood. For the objective of this project, not all categories are important individually, so related categories are clubbed together into some broad category resulting in only 10 categories. Then one hot encoding is applied on the “Category” column to get the dummies indicating which category a venue belongs to. This one hot encoded dataset has 9900 venues mapped to unique zip and neighborhood, which is also later used to obtain the frequency of each type of venue in a neighborhood.

3. Methodology

The first requirement it to get the dataset of all the zip codes that belong to the State of Illinois, USA. This dataset is available at the source mentioned in Section 2.1.1.1. After downloading the dataset as a .csv file, it can be opened in a jupyter notebook. This dataset has all the zip codes, corresponding coordinates, city and some other features that we don't need. Next requirement is of data about neighbourhoods in the city of Chicago, Illinois-USA. This data can be obtained using web scraping techniques from the source mentioned in Section 2.1.1.2.
We use Python requests library and beautifulsoup packages to extract the list of neighbourhoods data. Having obtained both these datasets, we need to merge them and extract common entries present in both of the data sources. This will give us all the zip codes, coordinates and neighborhood names belonging to the city of Chicago. After gathering the data, we will populate the data
into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This is necessarily a sanity check to make sure that the geographical coordinates we have indeed belong to the city of Chicago.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 5 kilometers. The Foursquare developer services allow for an app with a unique ID and Secret Key, using which their database can be queried for location data if provided with certain information about the location. API calls to Foursquare return the venue data in JSON format from which venue name, venue category, venue coordinates etc. can be extracted. This data is then analyzed to check how many venues were returned for each neighborhood, what all unique type of venues were reported. Since we get around 250 unique categories of venues spanning over a total of 9900 venues (100 for each neighborhood), they are a bit difficult to analyze, especially with the objective in mind. So, saveral relatable categories are clubbed together to reduce the information detail a little. Finally, the dataset has total 9900 venues belonging to 10 categories. We obtain the dummies corresponding to each category and then analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. We also use the "elbow" method on the full one hot encoded dataset to find the optimal number of clusters for these venues (the clustering method used is the Kmeans algorithm from "sklearn library"). This optimal number for our case is 5. After obtaining the optimal number of clusters, we use the venue frequency data for each neighborhood, and compress all the information to 2 dimensions using Principal Component Analysis (PCA). This is done maily for visualization purposes. We now apply clustering (with nclusters=5 obtained on one hot encoded dataset) on this dataset and plot the clusters. The clusters so obtained seem logical and not ambiguous enough to raise suspicion.

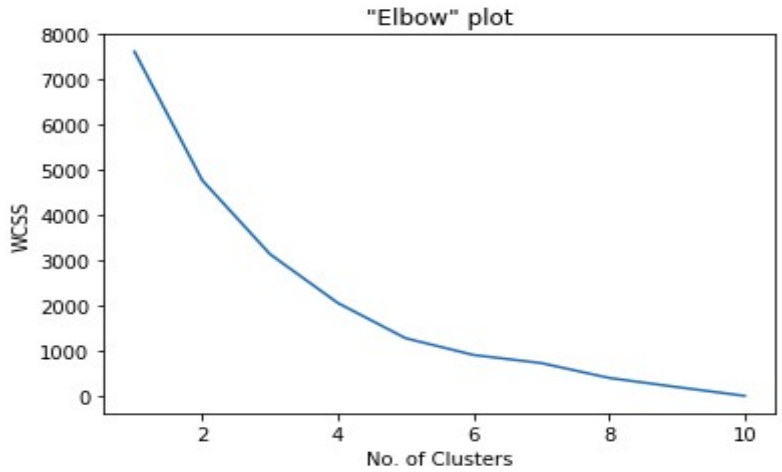
Having successfully performed clustering, we now focus on our main objective of location for the gym centre. We analyze the distribution of number of gyms present

in a neighborhood belonging to a particular cluster. This step reveals that the neighborhoods that belong to “Cluster 1” have comparatively larger number of

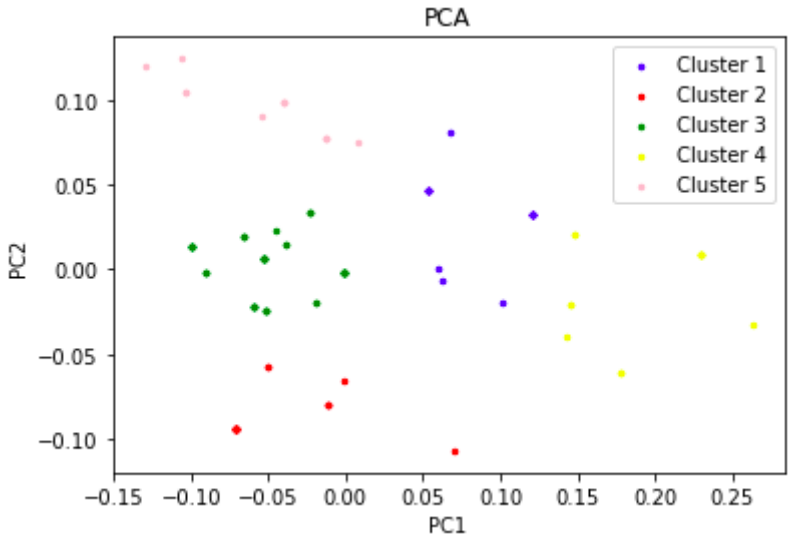
gyms/fitness centers. This indicates a high competition level, as well as bigger customer base.

Next, we extract the venues that are categorised as gym in the dataset and plot them on the map using their corresponding coordinates with the help of folium library.

We also report top 5 neighborhoods that belong to a particular cluster (in the order of cluster label) along with the corresponding number of gyms in each of those neighborhoods.



1. Elbow method to determine the optimal number of clusters



2. Reducing features through PCA, then clustering the frequency_of_venues data of neighborhoods into 5 clusters (as suggested by elbow method)

4. Results

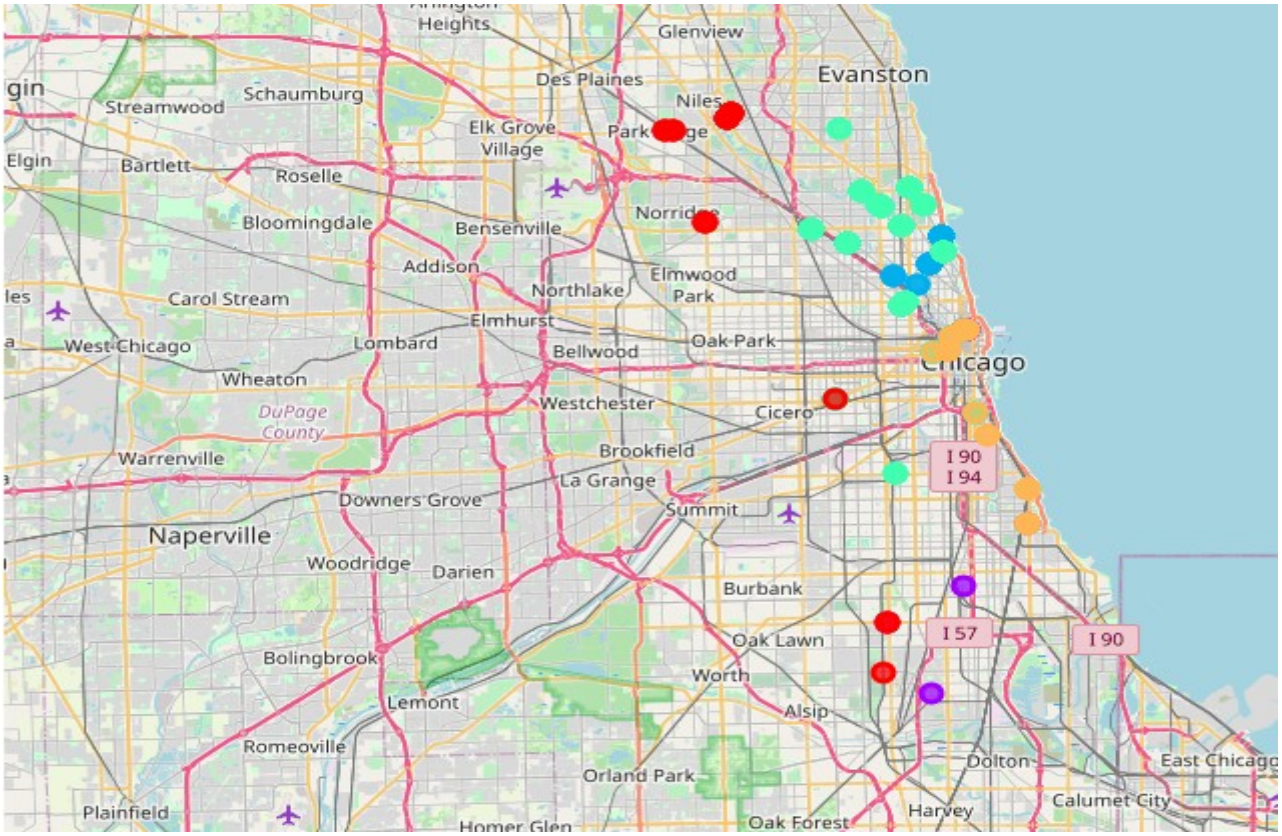
Following results are obtained in the analysis of the data while approaching the described business problem:

4.1. The neighborhoods of Chicago can be broadly grouped into 5 clusters based on 10 broad categories described in this project.

4.2. The distribution of number of gyms in the above clusters is different and can be described as:

- **4.2.1. Cluster 0:** neighborhoods have around 2-4 gyms on average (Medium)
- **4.2.2. Cluster 1:** neighborhoods have around 5 or more gyms on average (High)
- **4.2.3. Cluster 2:** neighborhoods have around 3 or more gyms on average (High)
- **4.2.4. Cluster 3:** neighborhoods have around 2-4 gyms on average (Medium)
- **4.2.5. Cluster 4:** neighborhoods have around 1-3 gyms on average (Low)

Blue, purple, green, orange and red colors represent gyms belonging to Cluster 0, 1 2, 3 and 4 respectively.



5. Discussion & Observations

5.1. On the map, we can see that more and more gyms are located in the northern parts of the city.
5.2. Cluster 1 and Cluster 2 (which are the clusters where more than 50% of neighborhoods having more than 6 gyms and 3 gyms respectively) also fall mostly towards the northern parts, indicating a higer level of awareness among people towards health as well as higher competition in gym business in these areas.

6. Conclusion

1. Aggressive strategy:

For clients who can boast of their highest quality gym centres, or clients who are capable enough (big "Fitness Chains") and favour aggressive strategy of dominating the competition should set up their gym centre(s) closer to the places (listed above) that are grouped under Cluster 1 or Cluster 2.

2. Medium-competition, medium profit strategy:

For clients who favour the "just right" competition, or are new and looking to establish their brand name in the market without taking extreme risks, and can cater to medium to small crowds qualitatively, it is recommended to set up new gym centers closer to the existing gym centres in the neighborhoods that belong to Cluster 0 or Cluster 3.

3. Minimum competition, Satisfactory profits:

For risk averse clients and business with comparatively small assets, who can not cater to large number of customers while maintaining a good quality of service, should set up their gym centres in the neighborhoods that belong to Cluster 4, preferably with a sufficient distance from the existing gym centres in those neighborhoods. This type of neighborhood is also suitable for clients who just want to do a market testing through a small project, or for clients who want to establish gym centre as a side business and can not devote much time for the growth of business.